PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies

PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies

Manuela Sanguinetti[⋆], **Cristina Bosco**[⋆], **Alberto Lavelli**[•], **Alessandro Mazzei**[⋆], **Oronzo Antonelli**[‡], **Fabio Tamburini**[◇]
[⋆]Department of Computer Science, University of Turin, Italy
[•]FBK-ict, Trento, Italy
[‡]DISI / [◇]FICLIT, University of Bologna, Italy
{msanguin, bosco, mazzei}@di.unito.it, lavelli@fbk.eu, antonelli.oronzo@gmail.com, fabio.tamburini@unibo.it

Abstract

Due to the spread of social media-based applications and the challenges posed by the treatment of social media texts in NLP tools, tailored approaches and *ad hoc* resources are required to provide the proper coverage of specific linguistic phenomena. Various attempts to produce this kind of specialized resources and tools are described in literature. However, most of these attempts mainly focus on PoS-tagged corpora and only a few of them deal with syntactic annotation. This is particularly true for the Italian language, for which such a resource is currently missing. We thus propose the development of PoSTWITA-UD, a collection of tweets annotated according to a well-known dependency-based annotation format: the Universal Dependencies. The goal of this work is manifold, and it mainly consists in creating a resource that, especially for Italian, can be exploited for the training of NLP systems so as to enhance their performance on social media texts. In this paper we focus on the current state of the resource.

Keywords: social media language, Twitter, Italian, Universal Dependencies

1. Introduction

The increasing reliance on the popularity of the Internet and social media in every-day life has led, among other things, to the proliferation of the so-called user-generated contents. As one of the most popular social media, Twitter is among the main providers of this type of contents, the usage of which in scientific research ranges from data analysis, sentiment analysis and opinion mining, to language technologies. Often, though, Twitter user-generated contents are not edited and/or revised for grammatical accuracy (Lynn et al., 2015). Furthermore, the limited number of characters for each tweet can stimulate creativity and encourage an innovative and non-standard usage of language conventions. As regards NLP, dealing with this kind of linguistic data presents a series of challenges, which are reflected in the lower output quality of various automatic tools and at different linguistic levels (see, e.g., Gimpel et al.(2011), Foster et al. (2011) and Ritter et al. (2011)).

These considerations highlight the need for properly annotated resources to provide adequate coverage of such a phenomenon. This is especially true considering the (relatively) little progress made in this field: for the Italian language in particular, at the time of writing, the only linguistically-annotated social media corpora we are aware of are those of Bosco et al. (2016) (i.e. the PoSTWITA corpus) and Rei et al. (2016), both annotated at PoS level only.

The contribution of the work hereby presented aims at filling this gap, by creating a treebank of Italian non-canonical texts retrieved from Twitter. The treebank has been built using as starting data the PoSTWITA corpus (mentioned above), and it has been syntactically annotated in compliance with the Universal Dependencies format.

Several goals motivate this work, among these: *a*) to provide a resource that can be used for parser training on standard (Bosco et al., 2008; Bosco et al., 2010; Bosco and Mazzei, 2013) and non-standard texts (whose results, in turn, can be exploited in sentiment analysis applications[1]), as well as for systematic linguistic analysis related to social media language (similar to what proposed in Hu et al. (2013)); *b*) in a long-term perspective, to encourage the creation of similar resources in languages other than Italian, supported by the availability of a shared representation format, namely the Universal Dependencies (possibly extended to cover social media linguistic phenomena).

As for the second point, that is the choice of the annotation format, Universal Dependencies is a recent project that has gained broad consensus over the last few years, becoming the reference framework for dependency annotation. Despite the critical points raised, for example, on some annotation choices (Gerdes and Kahane, 2016), or on the cross-resource consistency problem (de Marneffe et al., 2017), an ever increasing number of languages and resources have been (and are about to be) made available in this format; also two CoNLL Shared Task have been organized in 2017 and 2018, using UD treebanks as datasets. This highlights the need for a widely recognized standard to refer to, either in the process of creating a resource (from scratch or by conversion) or in the evaluation of NLP tools whose training is based on such a resource. Finally, one more factor that made us lean towards using the UD format is the possibility to extend the basic labels with subtypes that are useful for the representation of specific phenomena and are based both on the language at issue and, as in our case, on the peculiar linguistic features of the text type.

The paper focuses on the current state of the resource and is organized as follows. A brief survey is presented about related work in the next section. Section 3 describes the resource considering the conversion steps into UD format.

[1]In this sense, the creation of this resource is linked to the projects coordinated by the Computer Science Department of the University of Turin for the creation of automatic systems for online hate speech identification, in particular on Twitter. See: `http://hatespeech.di.unito.it/`

Section 4. focuses on the syntactic layer in particular, showing the main phenomena we dealt with for what concerns manual annotation. Finally, Section 5. describes the experiments we carried out by training and testing state-of-the-art parsers on the novel resource, while Section 6. closes the paper with few remarks on how we intend to follow up on this work.

2. Related Work

Social media texts fall under the broader language variety often referred to as non-canonical, or non-standard, language; its automatic processing and analysis is challenged namely by all those linguistic phenomena that deviate from what is conventionally conceived as the "norm", i.e. the standard language. In this section, we mention some of the attempts made in other related resources to tackle these challenges, especially as regards syntactic annotation.

Tweebank (Kong et al., 2014) is a corpus that presents a simplified, though linguistically-grounded, dependency-based scheme; the resource consists of unlabeled dependency graphs that allow multiple roots in case a tweet contains more than one utterance, and where just nodes with a syntactic function are explicitly selected.

Another attempt to properly annotate Web data was made in the English Web Treebank (Silveira et al., 2014), a collection of more than 16k sentences taken from various media, also available in UD format. In this resource, the treatment of Internet-related phenomena mainly entailed the revision of the inventory of dependency relations; in particular, new labels were introduced, that since then became an integral part of the UD scheme[2].

Other examples of non-canonical texts annotated in compliance with UD specifications are the Treebank of Learner English (Berzak et al., 2016) and the Singlish treebank (Wang et al., 2017). The first one is a collection of English as a Second Language (ESL) sentences, which thus contains a large number of non-standard syntactic structures due to grammatical errors made by the non-native English speakers. As regards their annotation, the main guiding principle prescribes to follow the *literal meaning*, emphasizing a syntactic analysis that is more faithful to the observed language usage. This is reflected, for example, in the annotation of a direct object as a non-core predicate dependent, if (wrongly) preceded by a preposition, or conversely, a non-core dependent annotated as predicate argument because of an elided preposition. The second example of UD format applied to non-standard texts is the one presented in Wang et al. (2017) and regarding the syntactically-annotated resource of Colloquial Singapore English (or Singlish)[3], an English-based creole language, frequently used in written forms of Web media. Most of the problems encountered in the annotation of such texts had to do with the treatment of terms and expressions imported from local languages, whose annotation is mainly based on the conventions of such languages rather than English, as well as on topic-prominence phenomena, copula and NP deletions, and inversions, all linguistic constructions that eventually have been modeled successfully with UD representation.

To conclude, we mention here the work presented by Martinéz Alonso et al. (2016), on the creation of a French corpus of user-generated (UGC) content with automatic PoS tagging and an experimental syntactic annotation (on a smaller sub-section of the corpus) using UD. Besides confirming, once again, the challenges posed to the treatment of UGC-related phenomena, the study also brings to light some critical points of UD format and specifications when it comes to deal with such issues, with an eye in particular on the tokenization, and the consequent syntactic annotation, of non-standard conflated tokens, as well as elliptical structures and disfluencies resulting from the time and space limitations posed by the medium used.

In Section 4., we describe our approach to such phenomena.

3. Introducing PoSTWITA-UD: Conversion and Current Dataset

PoSTWITA-UD has been created by enriching the dataset used for the EVALITA 2016 task of Part-of-Speech tagging of Social Media (Bosco et al., 2016). The original corpus consists of 6,438 tweets in the development set (114,967 tokens) and 300 tweets in the test set (4,759 tokens), annotated at PoS level only. The format of the resource, also shown in Figure 1, appears as a two-column text file with tweets identified by their IDs (in the header) and separated by blank lines; each word in the tweet has its own line, which in turn contains two tab-separated fields, for the word form and its Part of Speech respectively.

```
_____579013335921885184_____
@LudovicaCagnino    MENTION
Grazie  INTJ
AMORE   NOUN
```

Figure 1: Example of PoSTWITA original format.

The corpus was automatically tokenized with an adapted version of the Tweet-NLP tokenizer(Gimpel et al., 2011), PoS-tagged with the TnT tagger (Brants, 2000) trained on the UD_Italian treebank v1.3 (Bosco et al., 2013), and then manually corrected.

The whole process of conversion into UD and its annotation has been described in Sanguinetti et al. (2017), but for the sake of clarity, we summarize here the main steps we followed in order to get a fully UD-compliant resource.

Tokenization: no particular changes have been made in this sense from PoSTWITA to PoSTWITA-UD, except for preposition-article and verb-clitic contractions, that were left as single tokens in PoSTWITA and then splitted into the corresponding syntactic words during conversion into UD. All other tokenization choices remained unchanged in PoSTWITA-UD; this also entailed the occurrence of cases where multiple tokens were kept as a single one, whether mistakenly or on purpose, i.e. either because some typo occurred (e.g. "*anchio*" instead of "*anch'io*", "me too") or in

[2]These are `discourse`, `goeswith`, `list` and `vocative`.

[3]The resource is not available in the official UD repository, but here: `https://github.com/wanghm92/Sing_Par/tree/ud_tf0.12/Singlish/treebank`

	PoSTWITA	PoSTWITA-UD v2.1	PoSTWITA-UD v2.2
Annotation Layers	Part of Speech	Lemma Language-specific Tag (xpos) Morphological Features Syntactic Relation	
# of tweets	6,738	3,510	6,712
# of (syntactic) words	119,726	64,536	124,410

Table 1: Treebank basic statistics and differences between the original PoSTWITA corpus, used in the EVALITA campaign in 2016, and the converted versions in UD format: the one released in November 2017 (v2.1), containing just the first half of the corrected dataset, and the v2.2 that finally comprises the complete dataset. The overall number of tweets in PoSTWITA-UD v2.2 differs from the one in the original PoSTWITA due to the removal of duplicate tweets, while the change in the number of syntactic words also depends on the tokenization steps carried out during conversion (see Section 3.).

case of abbreviations of two or more words (e.g. "TT", that stands for "trend topic"), or even because of expressive intents (e.g. "*éStataPremiataUna*", "itHasBeenAwardedA"). Such cases, however, are not particularly frequent in our corpus, neither systematic. Considering that UD do not force the splitting of such conflated tokens, we decided to keep them unchanged (see Section 4.2. for their syntactic treatment).

PoS tagging: the original PoSTWITA corpus already contained UD PoS tags; however the standard UD tagset was extended by adding *a*) two labels for the contracted forms mentioned above (i.e. `ADP_A` and `VERB_CLIT` respectively), *b*) a number of other new labels for non-standard elements typically found in tweets, such as URLs (`URL`), email addresses (`EMAIL`), pictograms (`EMO`), hashtags (`HASHTAG`) and mentions (`MENTION`).
In PoSTWITA-UD, `ADP_A` and `VERB_CLIT` were completely removed, because of the splitting of such contractions, while the other Internet-specific tags all conflated into `SYM`, the tag used for symbols. Any other unconventional token whose tagging was not possible for some reason was assigned a `X` tag.

Lemmatization and morphological analysis: lemmas and mophological features were retrieved using AnIta (Tamburini and Melandri, 2012). However, as expected, the corpus also contains a whole host of non-standard word forms that were not recognized by the lemmatizer. In the spirit of leaving the texts as much intact as possible, we decided not to normalize such forms, which still appear in the resource as they do in the original tweet. On the other hand, in case of abbreviations (*ke*⇒ *che*, 'that'), word lengthening (*pizzaaaaaaa*⇒pizza), capitalization (*GOVERNO*⇒*governo*, 'government'), minor typos and grammatical errors (*anno* instead of *hanno*, 3rd pers.plur of *avere*, 'to have'), we manually inserted the lemma of their standard counterpart; for other out-of-vocabulary words, such as dialectal and foreign terms or unintelligible forms, the lemma remained the same as the word form. Note also that for abbreviations of multiple words we kept the abbreviation in the lemma field as well.

Syntactic analysis: this step has been performed first automatically, by training three parsers on Italian standard texts, namely those included in UD_Italian v2. The tools used were the graph-based (Bohnet, 2010) and transition-based (Bohnet and Nivre, 2012; Bohnet and Kuhn, 2012) MATE parsers, and RBG (Lei et al., 2014; Zhang et al., 2014b; Zhang et al., 2014a). The parsing step was performed on the entire resource and relied on the previously-annotated layers.
A first set of 300 tweets was then revised by two independent annotators in order to calculate the inter-annotator agreement (a Cohen's k = 0.92) and to test the parsers results.
Finally, the output that gained the best results (i.e. the one from the transition-based MATE parser) was chosen as the final version, and the two annotators completely revised it.
The first part (about 3,500 tweets) of the manually-corrected corpus was made available in November 2017, as part of the v2.1 release of Universal Dependencies. In view of the 2.2 version, whose official release is scheduled for April 2018, and of the upcomping CoNLL shared task, the second half as well is expected to appear in the resource, thus completing the whole annotation process.
To conclude this introductory section, Table 1 summarizes the treebank basic statistics, also highlighting the differences among the various versions of the resource.
The next section describes the guiding principles we adopted to annotate the treebank.

4. A UD-based Analysis of Italian Tweets: Updated Guidelines

As already described in Sanguinetti et al. (2017), the first tentative guidelines were drafted in parallel with an initial annotation experiment of 300 tweets. Since then, those guidelines have been further revised and integrated, in order to cover a wider range of UGC-related and other tricky phenomena.
The main guiding principle followed during the whole annotation process provides that what is understandable by a human should be annotated accordingly: this means that even in the presence of non-canonical tokens or structures, whenever the annotator is able to grasp their meaning with a certain degree of confidence, he/she is expected to encode it properly, according to such interpretation. On the other hand, while this served as a general guideline, more peculiar issues have also been encountered, the treatment of which required specific solving strategies. We grouped such issues into the following categories:

- meta-language tokens

- non-standard word forms
- juxtaposition and sentence linking
- elliptical structures

and we attempted to deal with them in a more systematic way.

4.1. Meta-Language Tokens

We include in this category the non-conventional textual elements typical of Twitter (as well as other social platforms), especially hashtags, mentions, pictograms, RTs (i.e. the tokens that usually precede a retweet), URLs and similar. These elements, if not inherent to the syntactic structure of the sentence, have been annotated using different criteria, depending on their type and their distinguishing features, also by introducing new specific label extensions.
More specifically, hashtags are considered as juxtaposed elements, and thus annotated using the `parataxis:hashtag` relation; mentions are annotated with the `vocative:mention` label while emoticons and emojis are treated as discourse markers, therefore they all bear the `discourse:emo` relation. Finally, the `dep` relation has been systematically used with URLs that are just appended at the end of a tweet, as well as with RT tokens. Figure 2 shows an example for each case:

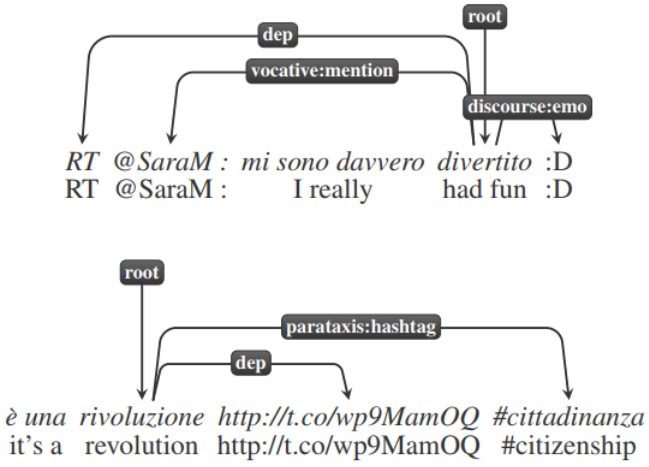

Figure 2: Examples of relations for meta-language tokens in PoSTWTITA-UD.

On the other hand, if syntactically integrated within the sentence, these same elements are annotated taking into account their actual syntactic role. In the sentence:

@AttilioFontains ha risposto al mio messaggio
(@AttilioFontains replied to my message)

the mention is considered as the subject of the sentence.

4.2. Non-Standard Word Forms

This category includes a wide range of different examples of various nature, from foreign and dialectal terms to mistakenly conflated forms (*cosè*, instead of *cos'è*, 'what is'), but also truncated words (due to space constraints) and completely unintelligible forms.
As also stated in Section 3., such forms were all PoS-tagged as `X` elements; still, the identification of their syntactic role remained quite unclear. A further distinction has thus been made, and each distinct case has been treated differently.

Code switching: if a single foreign or dialectal word occurs within the sentences, it is considered for its actual syntactic role, as also prescribed in similar cases for meta-language tokens (Section 4.1.); however, a phrase involving more than one token is considered as a flat structure[4] specified by the `:foreign` subtype.

Conflations: conflated forms are treated by assigning the syntactic relation associated with the word in the token that is promoted as head. In the sentence:

Fedenon ha proferito parola
(Fede-did-not say a word)

the token *Fedenon* is the concatenation of the proper noun *Fede* (abbreviation of *Federica*), which in this context is the subject of the verb *proferito* ('said')), and *non* ('not'), an adverbial modifier. Being a core argument, the proper noun is thus promoted as the head word and the whole token is annotated as `nsbuj`.

Truncated and unknown words: we also found a number of cases where the last word in the tweet was cut off, because of the character limits posed to tweets; even in these cases, we annotate the word according to its supposed syntactic role, whenever possible; if not, and, more generally, in case of unintelligible word forms, a `dep` relation is used.

4.3. Juxtaposition and Sentence Linking

From a syntactic point of view, UGC is also characterized by an abundance of paratactic and juxtaposed sentences; this is particularly true for Twitter posts, that often consist of more than one sentence, as below:

#michelebravi un nome una garanzia? Vediamo. Anzi sentiamo
(#michelebravi a name a guarantee? We'll see. Or rather, we'll hear)

Note also that, given the intended use of the resource as training set for NLP tools in a real-world setting, the main segmentation unit in PoSTWITA-UD is the complete tweet, rather than the single sentence. However, establishing a dependency link among such sentences is not a trivial task, especially considering the single-root constraint posed by UD scheme. Therefore, we systematically resorted to `parataxis` even to represent such inter-sentential links, although aware of the theoretical limits and the risk of an "overuse" of the label. The tweet above would thus be annotated as shown in Figure 3:
Furthermore, besides extending, in a sense, the applicability of this label to such a wider range of cases, we have also

[4] See `http://universaldependencies.org/u/dep/flat.html`

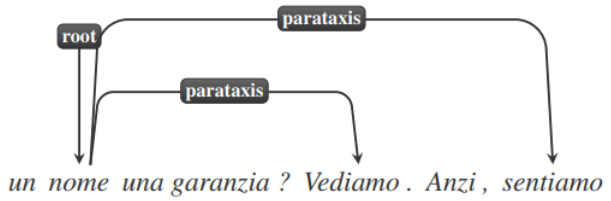

Figure 3: Example of `parataxis` used as inter-sentential link.

introduced a further distinction in its use with respect to other ones. More specifically, we have included new label extensions for five cases in particular (in addition to the one mentioned above for hashtags), even by partly mutuating the dependency labels from already existing treebanks[5]. Such cases comprise the dependency link used to identify appositive sentences (`parataxis:appos`), the semantically-void clauses used as discourse markers (`parataxis:discourse`), the parenthetical clauses that cannot be considered independent from the governing predicate (`parataxis:insert`), and the paratactic sentences having an implicit argumental role with respect to the governing predicate (`parataxis:nsubj` and `parataxis:obj`). All these relations are shown with practical examples in Figure 4.

4.4. Elliptical Structures

The need for immediacy in computer-mediated communication, its interactive nature, as well as the space constraints posed by the medium used, often result in a fragmentary writing (Martínez-Alonso et al., 2016) and very concise utterances, where more or less meaningful portions of a sentence are omitted[6]. As a consequence, and in a way that recalls the so-called *headlinese*, a given sentence may have function words removed, such as:

Manovra Governo Monti
((The) budget measures (of) Monti administration)

Given the preference of UD scheme in assigning headedness to content words, no solving strategy has been necessary for such cases. Even when copulas in copulative sentences were omitted, the main constituents of the sentence (i.e. the nominal predicate and its subject, if present) preserved their function, and the missing copula has been simply ignored.
However, elliptical structures can also reach a higher degree of complexity; in such cases, or at least whenever possible, we followed the main guideline, by attempting to interpret the missing context and to annotate the tweet accordingly. For example, in the sentence:

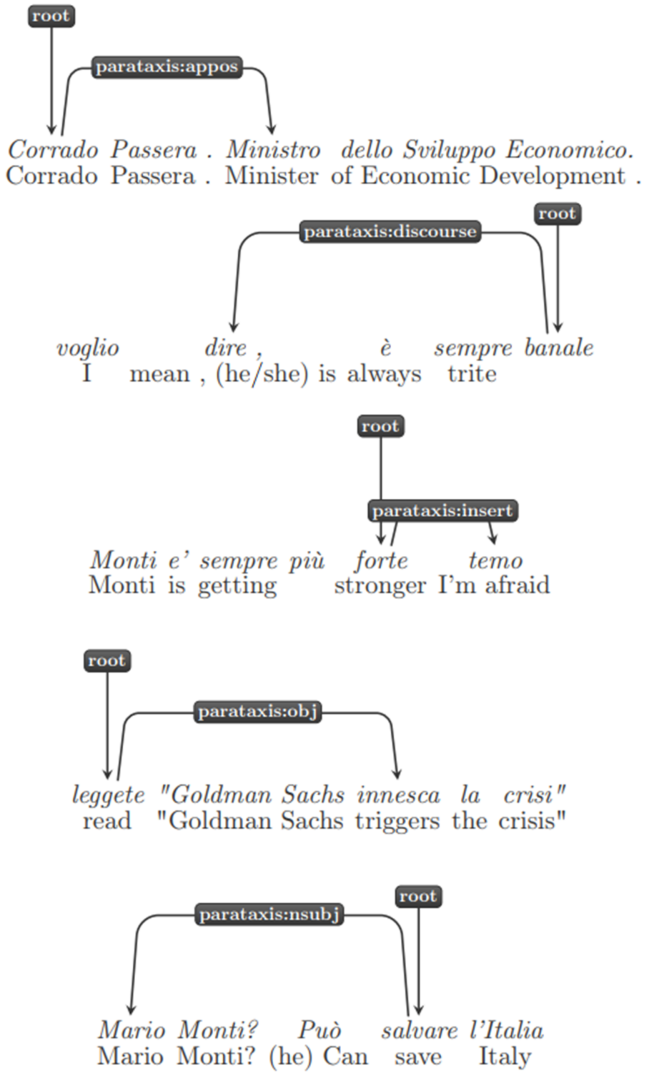

Figure 4: Examples of newly-introduced label extensions for paratactic structures in PoSTWTITA-UD.

Innalzata età minima donne 62 e uomini 66 dal 2018
(Raised (the) retirement age (of) women (to) 62 and (of) men (to) 66 (starting) from 2018)

besides the function words removal, a complex predicate ellipsis also occurs, which requires the special relation `orphan`, applied as shown in Figure 5.
In case the degree of uncertainty is such that it made the interpretation effort completely pointless, we rather linked the disconnected fragments, again, either with a `parataxis` or with an even more generic `dep` relation, depending on the tweet context.

5. Some Parsing Experiments

In order to test the effectiveness of the proposed annotations and to gather information about the difficulties in parsing tweet texts, we carried out some parsing experiments using state-of-the-art parsers available to the community. We

[5]Namely the UD_French-Spoken (Gerdes and Kahane, 2017) and UD_Slovenian-SST (Dobrovoljc and Nivre, 2016)

[6]In this respect, it is also similar to sign languages, where the lack of determiners and prepositions can be compensated with a sort of visual/spatial organization of the sentence (Mazzei et al., 2013).

Parser	UD_It		UD_PoSTW		UD_It+PoSTW	
	UAS	LAS	UAS	LAS	UAS	LAS
(Chen and Manning, 2014)	72.25%	63.98%	81.75%	76.34%	82.94%	77.60%
(Ballesteros et al., 2015)	73.98%	65.71%	84.28%	78.97%	85.21%	79.93%
(Kiperwasser and Goldberg, 2016): Transition	77.17%	68.12%	77.46%	68.95%	80.79%	73.36%
(Kiperwasser and Goldberg, 2016): Graph	75.96%	67.54%	79.49%	71.19%	81.43%	73.49%
(Andor et al., 2016)	65.72%	52.31%	77.88%	67.11%	79.52%	69.04%
(Cheng et al., 2016)	76.94%	67.54%	86.12%	79.89%	86.85%	80.93%
(Dozat and Manning, 2017)	**77.48%**	**68.22%**	**86.38%**	**80.53%**	**86.95%**	**81.49%**
(Shi et al., 2017a; Shi et al., 2017b)	71.69%	66.89%	81.41%	74.73%	83.48%	76.54%
(Nguyen et al., 2017)	70.84%	61.21%	83.37%	76.95%	84.03%	77.98%

Table 2: Evaluation results on the PoSTWITA 2.2 test set using the different setups: training using only UD_Italian 2.1 (UD_It), training using only PoSTWITA 2.2 (UD_PoSTW) and training using both resources (UD_It+PoSTW). All the parser outputs were evaluated by using the standard script devised for the CoNLL-X evaluation.

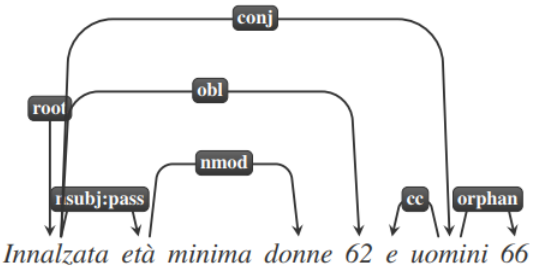

Figure 5: Annotation example of complex predicate ellipsis.

organized the experiments into three different setups:

1. in the first setup we trained and tuned the parsers by using the train and development sets belonging to the general Italian UD treebank UD_Italian v2.1 (UD_It). This setup is useful for comparing the parsers' performance when trained with general texts (out-of-domain) with the results obtained on Twitter data, similarly to what already proposed in Sanguinetti et al. (2017);
2. considering the importance of training with in-domain data, especially when it comes to non-standard texts (and UGC in particular), in the second setup we trained and tuned the parsers by using the UD_PoSTWITA v2.2 only (UD_PoSTW);
3. in the last set of experiments we trained and tuned the parsers by using both resources (UD_It+PoSTW).

In all the experiments the parsers were tested on the UD_PoSTWITA v2.2 test set. Table 2 shows the results for all setups. An in-depth analysis of parsing problems for tweets is well beyond the scope of this paper, but we can draw some provisional observations by examining the obtained results: first of all, the parser from Dozat and Manning (2017) consistently outperforms all the other competitors exhibiting a nice robustness also for this kind of texts. Second, even if the data sets from UD_PoSTWITA v2.2 are smaller than UD_Italian v2.1, in-domain data are fundamental for getting reliable results. Third, adding the UD_Italian 2.1 treebank does increase the performance, but only to a limited extent, suggesting, again, that out-of-domain data are less useful for obtaining good results.

Finally, considering the use of *parataxis* also for sentence linking (which is an unusual application of this label, as explained in Section 4.3.), we observed the performance of Dozat and Manning parser, trained using the UD_It+PoSTW setup, on this relation. The results are reported in Table 3 and show that, probably due to such specific use of the label, the parser results on its proper annotation are (relatively) low. We finally compared LAS and UAS scores on `parataxis` subtypes, observing that, except for hashtags, they were very poorly annotated, mainly because of their far lower frequency also in the training set.

Relation	Tot.	LAS	UAS
parataxis	509	60.31	66.60
parataxis:appos	16	12.50	75.00
parataxis:discourse	6	0	0
parataxis:hashtag	216	61.11	83.33
parataxis:insert	4	25	50.00
parataxis:nsubj	3	0	66.67
parataxis:obj	13	0	67.93

Table 3: Evaluation results on the PoSTWITA 2.2 test set for the specific *parataxis* relation and its subtypes, using the UD_It+PoSTW setup.

As already stated above, a more in-depth analysis would be necessary to study and better understand the main challenges of automatic tools in providing an accurate analysis of social media texts, specifically for Italian. We leave this point to future work.

6. Conclusion

In this paper, we presented the development of an Italian social media corpus annotated according to Universal Dependencies, and included in the official UD repositories since v2.1. This work aimed at showing how a *de facto* standard such as UD can be extended and applied also to one of the currently more widespread types of non-standard texts. Moreover, provided that the resource is especially tailored for NLP tools training, we also proposed a preliminary set of experiments with state-of-the-art parsing systems, in order to pave the way for an in-depth error analysis that takes into account all the annotation issues discussed in Section

4. in a more systematic way, thus overcoming the limitations of the present work. In the next future, we intend to further investigate this line of research also testing this approach on new Twitter data; as a side effect, this will result in a richer resource for training purposes.

Acknowledgements

The work of Cristina Bosco and Manuela Sanguinetti has been partially funded by Fondazione CRT (*Hate Speech and Social Media*, project n. 2016.0688) and by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, project S1618_L2_BOSC_01).

7. References

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany, August. Association for Computational Linguistics.

Ballesteros, M., Dyer, C., and Smith, N. A. (2015). Improved transition-based parsing by modeling characters instead of words with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 349–359, Lisbon, Portugal, September. Association for Computational Linguistics.

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., and Katz, B. (2016). Universal dependencies for learner english. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany, August. Association for Computational Linguistics.

Bohnet, B. and Kuhn, J. (2012). The best of both worlds – a graph-based completion model for transition-based parsers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–87, Avignon, France. Association for Computational Linguistics.

Bohnet, B. and Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea. Association for Computational Linguistics.

Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China. Coling 2010 Organizing Committee.

Bosco, C. and Mazzei, A. (2013). The evalita dependency parsing task: From 2007 to 2011. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7689 LNAI:1–12.

Bosco, C., Mazzei, A., Lombardo, V., Attardi, G., Corazza, A., Lavelli, A., Lesmo, L., Satta, G., and Simi, M. (2008). Comparing italian parsers on a common treebank: The evalita experience. pages 2066–2073.

Bosco, C., Montemagni, S., Mazzei, A., Lombardo, V., Dell'Orletta, F., Lenci, A., Lesmo, L., Attardi, G., Simi, M., Lavelli, A., Hall, J., Nilsson, J., and Nivre, J. (2010). Comparing the influence of different treebank annotations on dependency parsing. pages 1794–1801. cited By 9.

Bosco, C., Montemagni, S., and Simi, M. (2013). Converting Italian treebanks: Towards an Italian Stanford Dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69.

Bosco, C., Tamburini, F., Bolioli, A., and Mazzei, A. (2016). Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian task. In *Proceedings of Evalita 2016*.

Brants, T. (2000). Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 224–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October. Association for Computational Linguistics.

Cheng, H., Fang, H., He, X., Gao, J., and Deng, L. (2016). Bi-directional attention with agreement for dependency parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2204–2214, Austin, Texas, November. Association for Computational Linguistics.

de Marneffe, M.-C., Grioni, M., Kanerva, J., and Ginter, F. (2017). Assessing the annotation consistency of the universal dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy*, pages 108–115.

Dobrovoljc, K. and Nivre, J. (2016). The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28*, pages 1566–1571. European Language Resources Association (ELRA).

Dozat, T. and Manning, C. D. (2017). Deep biaffine attention for neural dependency parsing. In *Proceedings of the 2017 International Conference on Learning Representations*.

Foster, J., Çetinoglu, Ö., Wagner, J., Roux, J. L., Hogan, S., Nivre, J., Hogan, D., and van Genabith, J. (2011). #hardtoparse: POS tagging and parsing the twitterverse. In *Analyzing Microtext, Papers from the 2011 AAAI Workshop, San Francisco, California, USA, August 8, 2011*.

Gerdes, K. and Kahane, S. (2016). Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. In *Proceedings of Linguistic*

Annotation Workshop (LAW).

Gerdes, K. and Kahane, S. (2017). Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral: le cas de la macrosyntaxe. In *Actes de l'atelier "ACor4French -– Les corpus annotés du français"*, pages 1–9, Orléans, France.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hu, Y., Talamadupula, K., and Kambhampati, S. (2013). *Dude, srsly?*: The surprisingly formal nature of twitter's language. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 244–253.

Kiperwasser, E. and Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (2014). A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar. Association for Computational Linguistics.

Lei, T., Xin, Y., Zhang, Y., Barzilay, R., and Jaakkola, T. (2014). Low-rank tensors for scoring dependency structures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1381–1391, Baltimore, Maryland. Association for Computational Linguistics.

Lynn, T., Scannell, K., and Maguire, E. (2015). Minority language twitter: Part-of-speech tagging and analysis of Irish tweets. In *Workshop on Noisy User-generated Text*, Beijing, China.

Martínez-Alonso, H., Seddah, D., and Sagot, B. (2016). From noisy questions to minecraft texts: Annotation challenges in extreme syntax scenario. In *Proceedings of the 2nd Workshop on Noisy User-generated Text, NUT@COLING 2016, Osaka, Japan, December 11, 2016*, pages 13–23.

Mazzei, A., Lesmo, L., Battaglino, C., Vendrame, M., and Bucciarelli, M. (2013). Deep natural language processing for italian sign language translation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8249 LNAI:193–204.

Nguyen, D. Q., Dras, M., and Johnson, M. (2017). A novel neural network model for joint pos tagging and graph-based dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 134–142, Vancouver, Canada, August. Association for Computational Linguistics.

Rei, L., Mladenić, D., and Krek, S. (2016). A multilingual social media linguistic corpus. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, Ljubljana, Slovenia.

Ritter, A., Mausam, S. C., and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, pages 1524––1534, Stroudsburg, PA, USA.

Sanguinetti, M., Bosco, C., Mazzei, A., Lavelli, A., and Tamburini, F. (2017). Annotating italian social media texts in universal dependencies. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy*, pages 229–239.

Shi, T., Huang, L., and Lee, L. (2017a). Fast(er) exact decoding and global training for transition-based dependency parsing via a minimal feature set. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 12–23, Copenhagen, Denmark, September. Association for Computational Linguistics.

Shi, T., Wu, F. G., Chen, X., and Cheng, Y. (2017b). Combining global models for parsing universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 31–39, Vancouver, Canada, August. Association for Computational Linguistics.

Silveira, N., Dozat, T., Marneffe, M.-C. D., Bowman, S., Connor, M., Bauer, J., and Manning, C. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tamburini, F. and Melandri, M. (2012). AnIta: a powerful morphological analyser for Italian. In *Proceedings of Language Resources and Evaluation Conference 2012*, pages 941–947.

Wang, H., Zhang, Y., Chan, G. L., Yang, J., and Chieu, H. L. (2017). Universal dependencies parsing for colloquial singaporean english. In Regina Barzilay et al., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1732–1744. Association for Computational Linguistics.

Zhang, Y., Lei, T., Barzilay, R., and Jaakkola, T. (2014a). Greed is good if randomized: New inference for dependency parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1013–1024, Doha, Qatar. Association for Computational Linguistics.

Zhang, Y., Lei, T., Barzilay, R., Jaakkola, T., and Globerson, A. (2014b). Steps to excellence: Simple inference with refined scoring of dependency trees. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Baltimore, Maryland. Association for Computational Linguistics.